宋词字-音转换中多音字处理研究

赖兴邦
（厦门大学智能科学系　福建 厦门　361005）

【摘　要】：字音转换问题一直是中文语音合成系统中不可缺少的模块，而确定多音字在文章的特定环境中读什么音是其核心问题。本文以全宋词为研究对象，实现了全宋词的字音转换，其中多音字消歧方法有三种：条件策略，词性策略，格律校正。本系统采用现代汉语字音转换方法和宋词格律特点结合来实现，其中宋词字音转换的多音字标注准确率提高到 96.89%

【关键词】：字音转换 多音字 宋词 格律 字符串匹配

0、引言

随着中文信息处理研究的深入，人们在关注现代汉语信息处理研究的同时，开始了古代汉语的研究。利用计算机技术对宋词字音转换后，更便于现代人感受宋词的声调铿锵、音律和谐。

字音转换在语音合成系统(Text-To-Speech)中任务就是将一系列文字系列转换成对应的拼音系列。字音转换中关键和难点就是如何解决一字多音问题，现代汉语主要采取有两种方法[1]：(1)基于规则体系的方法；(2)基于统计机器学习的方法。北大语言所在宋诗自动注音[2]中提到古汉语中经常使用单字词，词与词之间缺乏固定的结合关系，给自动注音带来很大的困难，提出利用宋诗自身音韵特点的规则策略。

基于以上的认识，本文通过一种新的结合方法来解决全宋词的字音转换。也希望能利用计算机寻找声韵流变的轨迹，解决声韵学的一些问题。

1、基于宋词语料库多音字的统计和分析

对 3000 首宋词手工注音结果进行计算机统计和分析发现以下几个现象：

1)大多数多音字也有现代汉语出现的高频音[3]。如："处"共出现 913 次，其中处(chu4)902 次，明显高于处(chu3) 11 次。

2)有些字有两种意义(往往词性也不同)，同时也有两种读音。例如"为"字，用作动词的时候解作"做"，就读平声(阳平)；用作介词的时候解作"因为"、"为了"，就读去声。在古代汉语里，这种情况比现代汉语多得多。

3)宋词作词一般是按照某种乐调曲拍之谱填制歌词。曲调的名称如《菩萨蛮》《蝶恋花》《念奴娇》等叫做"词调"或"词牌"，按照词调作词称为"倚声"或"填词"。词牌规定了词字数、平仄、押韵。

在此，本文提出先利用现代汉语使用的条件策略，词性策略，然后通过宋词格律校正来解决宋词多音字问题，最终取得满意的结果。下面详细介绍该系统的设计和实现。

2、系统的设计

2.1 系统的框架结构及流程

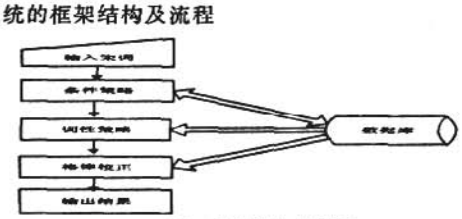

图 1 系统框架结构图

2、多音字注音策略

下面详细介绍本系统用到的三种多音字注音策略：

2.1 条件策略[4]

令多音字 Y 有 n 个音($X_1, X_2, \cdots, X_n$)，则 Y 读音为 X_i ($1 \leqslant i \leqslant n$) 的概率可表示如下：

$$P(X_i \mid Y) = \frac{P(X_i, Y)}{P(Y)} \approx \frac{F(X_i, Y)}{F(Y)}$$

$F(X_i, Y)$ 表示注音结果中字 Y 读音为 X_i 的次数，$F(Y)$ 表示语料库中字 Y 出现的次数。

首先在 *3000* 首手工注音熟语料库中算出所有多音字的参数 $P(X_i \mid Y)$，条件策略就是使得字 Y 的读音选为：$arg\ max_i P(X_i \mid Y)$。人工校对语料保存后将改变 $P(X_i \mid Y)$ 的值，也就意味着它是随语料库动态改变的。

2.2 词性策略

该方法针对一类特殊的多音字，这类多音字的每个词性对应于一个拼音(即词性对拼音的映射是一对一或多对一的)，如图(2)所示。只要确定了字的词性，那么它的读音也就确定了。

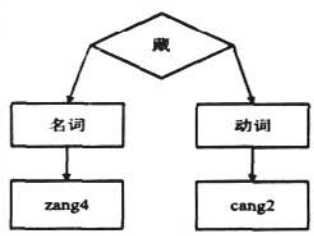

图 2 词性和读音的对应关系

本文数据库中收集了该类多音字 66 个，下面给出多音字"藏"的一个例子：

[芦/n 花/n]/XN [何/rd 处/n]/rd 藏/v 。/w

例子中的"藏"是动词，所以标注读音为"cang2"。词性策略对词性标注的准确率要求很高。本文用的分词和词形标注跟通用的不大一样，因为古汉语大多是单字成词，所以分词标注时尽可能标注每个字的词性，这样对词性策略有利。

2.3 格律校正

找到对应词谱，然后根据该词谱规定的多音字所在位置平仄和押韵，选择符合规则的读音。

字符表 T={○,●,△,▲,⊙,◎}，宋词格律和词体(词谱)都是由 T 中的字符表示成的字符串，宋词格律根据注音字库，把宋词转换成由 T 中字符表示的字符串(去除宋词中的标点符号)。但因存在同一词牌可能对应多个词体，到底选择哪个词体，这里定义相似度的概念，相似度反应宋词格律与词体的相似程度，具体定义为宋词格律字符串 S 转变为词体字符串 T 需要的字符编辑次数。字符编辑方法定义如下：

(1)用 T 中的一个符合 b 代替 S 中的一个符号 a

(2)插入 T 中的一个符号 a

(3)删除 S 中的一个符号 a

为了定义 S 和 T 之间的相似度，首先，将编辑操作和代价联系起来，表现为 $c(e_i)$ 其中，e_i 是任何一种编辑操作。然后，对于一个序列 s，它经过 k 个编辑操作将 S 转变为 T，总的代价值为：

$$c(s) = \sum_{i=1}^{k} c(e_i)$$

将 S 和 T 之间的相异度定义为将 S 转变为 T 的所有 n 个编辑操作序列中的最小代价：

$$d(s,t) = \min\{c(s_1), c(s_2), \ldots, c(s_n)\}$$

宋词格律 S 对应的词体就是所有 m 个词体与 X 相似度最大的词体，换言之也就是相异度最小的词体：

$$S(x) = \min\{d(s,t_1), d(s,t_2), \ldots, d(s,t_m)\}$$

计算 d(s,t)利用一种改变的 Levenshtein 距离算法，如图 3：

以矩阵的形式给出将 S 转变为 T 所有可能编辑操作系列，然后根据这个矩阵计算出最终结果。

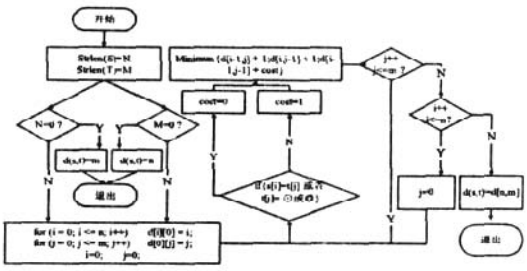

图 3 改变的 levenshtein 距离算法

3、字- 音转换步骤及实验结果分析

3.1 字- 音转换步骤

在此，以"迎客西来送客行。堆堆历历短长亭。殢人残酒不能醒。烟染暮山浮紫翠，霜凋秋叶复丹青。凭谁图写入银屏。"（作者：石孝友；词牌：浣溪沙）为例介绍本文的多音字部分的转换步骤：

从注音字库中可知该词中存在的多音字：迎(ying2/ying4)，行(xing2/hang2)，长(chang2/zhang3)，醒(xing3/xing1)，屏(ping2/bing1/bing3)

步骤 1：根据条件策略，多音字读音全部选择最高频音：

迎-{○}-ying2 行-{△}-xing2 长-{●}-zhang3

醒-{▲}-xing3 屏-{○}-ping2

步骤 2：机器自动切分及词性标注为"[迎/ v 客/n]/ VO [西/f 来/ v]/XV [送/ v 客/ n]/ VO 行/ v 。/w 堆堆/q 历历/d [短/a 长/a]/CA 亭/n 。/w 殢/ v 人/n [残/a 酒/n]/XN [不/d 能/ v]/XV 醒/ v 。/w 烟/n 染/ v [暮/t 山/n]/XN 浮/ v [紫/n 翠/n]/CN ，/w 霜/n 凋/ v [秋/t 叶/n]/XN 复/ v [丹/n 青/n]/CA 。/w 凭/ v 谁/r [图/ v 写/ v]/CV 入/ v [银/a 屏/n]/XN 。/w "

由词性策略能确定该类多音字是：行{动词- - xing2}，长{形容词- - chang2}则读音为：

迎-{○}-ying2 行-{△}-xing2 长-{○}-chang2

醒-{▲}-xing3 屏-{○}-ping2

步骤 3：把宋词转换为由字符表 T={○,●,△,▲,⊙,◎}的字符串，其中多音字按步骤 1、2 后的读音转换。接着选择词谱，由词谱数据库可得词牌《浣溪沙》共有 5 体，通过计算 5 个词体与该词格律相似度知，相似度最大的词体是："◎●⊙○◎●△◎○⊙●●○△⊙○◎●●○△⊙●◎○○●●◎○⊙●●○△◎○⊙●●○△"

由上面的词体可知：迎- -◎ 行 - -△ 长 - -○ 醒 - -△ 屏- -△

其中：行、醒、屏都是押"庚"平韵。

注意到经步骤 1、2 后，多音字"醒"不符合词谱，应校正为：醒-{△}-xing1 最终这首宋词字音转换为："ying2, ke4, xi1, lai2, song4, ke4, xing2, 。, dui1, dui1, li4, li4, duan3, chang2, ting2, 。ti4, ren2, can2, jiu3, bu4, neng2, xing1, 。yan1, ran3, mu4, shan1, fu2, zi3, cui4, , shuang1, diao1, qiu1, ye4, fu4, dan1, qing1, 。ping2, shui2, tu2, xie3, ru4, yin2, ping2, 。"

由上面的宋词举例转换注意到：

1、如果只使用策略 1，"长"高频音是"zhang3"，在这里正确读音是"chang2"。

2、如果只使用策略 2，只能确定该类多音字"行"，"长"。

3、如果只使用策略 3，"迎"对应的是"◎"，可平可仄，无法判断该字读音。使用策略 1，"迎"存在明显的高频音(ying2)，则可以正确标注为(ying2)。

4、"醒"在现代只有一个读音(xing3)，而在古代是多音字，也就是说古汉语跟现代现代汉语在读音上是存在差别的，而且差别非常大。

3.2 实验及结果分析

先从 3000 首手工标注的熟预料中提取多音字的发音频率，存入注音字库；接着对 3000 首（原 3000 首手工标注集外的）生语料宋词采用本文提到的 3 种策略结合的方法进行文语转换实验，经统计这 3000 首宋词总字数 305808 个字，其中多音字 47679 个，所以多音字占 15.6%

本文采用的评测标准是：

字音转换多音字的标注准确率=正确标注的多音字总字数/总共标注的多音字总字数

标注结果如表 1：

测试集（首）	策略1	组合方法
1000	92.81%	95.83%
2000	93.07%	96.31%
3000	93.90%	96.89%

表 1 多音字标注正确率

从实验数据对比可以看出，采用本文提到的三种策略结合起来的方法，能弥补彼此的不足，共同提高多音字标注准确率到 96.89%，整体宋词字音转换错误率降低到 4.8‰，因此说明该方法是有效的。但同时不足的地方是：

（一）宋词的自动分词以及词性标注的准确率还存在差距，这样必然影响到词性策略的作用效果。

（二）我们所谓的正确标注即人工标注，由于古今读音的差别以及人为理解的误差使得正确标注还有待改善。

4、结束语

本文是受现代汉语文语转换的启发，结合宋词固有的特性，尝试通过结合几种多音字注音的方法，对宋词进行文语转换，结果是令人满意的。

可以通过注音字典中的现代汉语读音信息与宋词词谱规则的对比分析来研究古今音的演变，本文对宋词的格律验证有一定的作用；同时也希望能对宋词研究提供一些启迪。

下一步要做的工作就是，统计每个多音字标注结果和标注准确率，然后研究标注准确率特别低的一些词语，找出问题所在，设法予以解决。另外还要研究的是如何提高宋词的词性标注准确率。

参考文献：

1.范明,胡国平,王仁华. 汉语字音转换中的多层面多音字读音消歧[A].计算机工程与应用,2006.02:167- 170

2.穗志方, 俞士汶, 罗凤珠. 宋代名家诗自动注音研究及系统实现[A]. 中文信息学报, 1997, 12(2): 44- 53

3.张子荣, 初敏.解决多音字字- 音转换的一种统计学习方法[A].中文信息学报, 2002, 03: 39- 45

4.潘慎. 词律辞典[M]. 山西:山西人民出版社,1982

5.钦定词谱[M]. 北京:北京人民出版社,1983

6.龙榆生. 唐宋词格律[M]. 上海:上海古籍出版社,1978

7.王兆鹏, 刘尊明. 宋词大辞典[M]. 南京: 凤凰出版社, 2003

8.陆辅之. 续修四库全书?词旨[M]. 上海:上海古籍出版社,1997

9.丁声树.古今字音对照手册[M]. 北京: 中华书局出版社,1981